Quantifying Uncertainty Through Coverage in Latent Semantic Space

Gabriele Carrino
IBM Research / Zurich
TU Wien / Vienna
gabriele.carrino@ibm.com

Brown Ebouky
IBM Research / Zurich
ETH / Zurich
brown.ebouky@ibm.com

Katja Hose
TU Wien / Vienna
katja.hose@tuwien.ac.at

Mattia Rigotti
IBM Research / Zurich
mrg@zurich.ibm.com

Abstract

Large language models often generate plausible yet unreliable answers, making uncertainty quantification essential for detecting potential hallucinations. Prior work has shown that the information needed to detect hallucinations and quantify a model's confidence in its generations may already be present in its internal states, but how to extract this information effectively remains an open problem. We introduce Semantic Coverage, a white-box, representation-level spectral uncertainty estimator inspired by coverage in reinforcement learning. Given multiple sampled generations for a prompt, Semantic Coverage quantifies uncertainty by measuring how broadly their internal representations span the model's latent semantic space. Across three open-weights language models and four short-form question-answering benchmarks, Semantic Coverage achieves the strongest overall performance among the evaluated methods and consistently outperforms prior representation-level methods in quantifying LLM uncertainty.

1 Introduction

Large Language Models (LLMs) have shown remarkable capabilities across a wide range of tasks, leading to their adoption in critical domains such as healthcare (He et al., 2025), finance (Nie et al., 2024), and law (Lai et al., 2023). However, so-called hallucinations remain a major limitation: LLMs can produce fluent and plausible outputs that are factually incorrect or misleading. Reliable deployment of LLM-powered systems therefore requires mechanisms for assessing when their outputs can be trusted (Amer-Yahia et al., 2025). This has uncertainty quantification a central research problem in the LLM literature (Farquhar et al., 2024; Kadavath et al., 2022; Chen et al., 2024; Shorinwa et al., 2025). Existing uncertainty estimates are often derived from prompt-based self-evaluation (Kadavath et al., 2022; Ren et al., 2023), token-level probabilities (Ren et al., 2022; Malinin and Gales, 2020; Dinh and Niehues, 2025), or semantic-level comparisons (Farquhar et al., 2024; Qiu and Miikkulainen, 2024). Self-evaluation methods are attractive because they require only black-box access, but they depend on the model's ability to express calibrated confidence (Xiong et al., 2024; Kapoor et al., 2025). Token-level methods are sensitive to surface-level variations that preserve the same semantic content. Semantic-level approaches operate at the level of meaning and are less affected by variation in surface form, but they typically rely on post-hoc semantic comparison after decoding, often through auxiliary models (Qiu and Miikkulainen, 2024; Farquhar et al., 2024). At the same time, recent work suggests that the information needed for hallucination detection may already be present in the model's internal representations (Ji et al., 2024; Chen et al., 2024), which have been shown to contain semantic information (Liu et al., 2024), signals predictive of truthfulness (Azaria and Mitchell, 2023), hallucinations (Su et al., 2024; Du et al., 2024), and uncertainty (Chen et al., 2024; Kossen et al., 2024). However, how to effectively extract this information remains an open problem. In this work, we approach this problem through the lens of continuous control and reinforcement learning, where exploration can be quantified by how broadly a trajectory spans the representation

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 2449–2465
November 6-10, 2026 ©2026 Association for Computational Linguistics

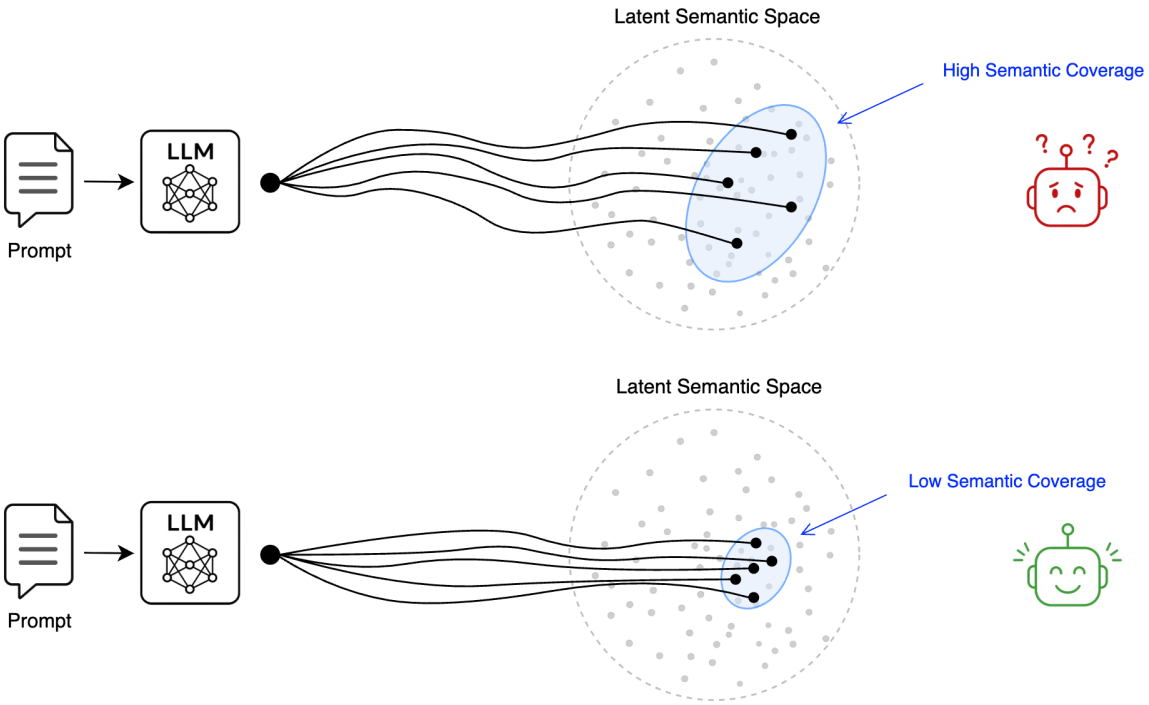


Figure 1: Conceptual overview of Semantic Coverage. When the model is confident, sampled internal representations are close in the latent semantic space. When it is uncertain, they span a broader region.

space (Wagenmaker et al., 2025). We adapt this methodology to Large Language Models by extracting semantic representations from the model's hidden states. When the model is confident, sampled response representations should remain concentrated; when it is uncertain, they should span a broader region of the latent semantic space. This motivates *Semantic Coverage*, a spectral uncertainty estimator that quantifies how broadly sampled representations span the latent semantic space.

Outline. The remainder of the paper is structured as follows. Section 2 reviews related work on LLM uncertainty quantification and introduces the representation-space coverage perspective that motivates our approach. Section 3 presents Semantic Coverage and its formulation. Section 4 describes the experimental setup and reports the main results and ablation studies. Section 5 concludes the paper, followed by a discussion of limitations and future directions.

Contributions. We introduce Semantic Coverage, a white-box uncertainty estimator computed from internal representations of multiple sampled generations for the same prompt. It quantifies how broadly sampled representations span the model's latent semantic space. Empirical results across three open-weights LLMs and four short-form QA benchmarks show that Semantic Coverage achieves the strongest overall performance among the evaluated methods and consistently outperforms prior representation-level methods.

2 Background and Related Work

Reinforcement Learning. Reinforcement learning (RL) studies how an agent learns through interactions with an environment. The sequence of states visited and actions taken by the agent is commonly represented as a trajectory, $\tau = (s_1, a_1, s_2, a_2, \ldots, s_t, a_t)$. A classical line of work drives exploration via uncertainty estimates over outcomes or value functions (Auer et al., 2002; Osband et al., 2016; Zhu and Rigotti, 2021). More recently, exploration has also been quantified through the coverage of the visited state history, $S = (s_1, s_2, \ldots, s_t)$, in a learned representation space (Wagenmaker et al., 2025). Under this view, an agent is more exploratory when the representations of the visited states span a broad region of the learned representation space, and less exploratory when they remain concentrated in a smaller region. We use this notion of representation-space coverage as the conceptual starting point for Semantic Coverage, while adapting it from trajectories of

environment states to sets of sampled generations.

Uncertainty Quantification. In this work, we study the problem of estimating how reliably an LLM can answer a given prompt. Given an input prompt x, the goal is to assign a scalar uncertainty score $u(x)$ that reflects the model's uncertainty about answering that prompt. Such a score can be used to assess whether the resulting answer should be trusted, for example by identifying potentially hallucinated outputs.

Related Work. Prior work on uncertainty quantification in LLMs can be grouped according to the signals they use. A first line of work relies on *self-evaluation* (Kadavath et al., 2022; Ren et al., 2023). A notable example is $\Pr(\text{True})$, proposed by Kadavath et al. (2022), where the model is asked to assess the correctness of its own answer. While simple, these methods rely on the model's ability to express calibrated uncertainty. Empirical studies show that such self-reported confidence is often overconfident and not well calibrated in open-ended generation, which limits its effectiveness as a standalone uncertainty estimate (Xiong et al., 2024; Kapoor et al., 2025). A second family estimates uncertainty from *token-level* probabilities. Straightforward choices include perplexity, length-normalized entropy (Ren et al., 2022; Malinin and Gales, 2020) and G-NLL, which uses the negative log-likelihood of the greedily decoded output (Aichberger et al., 2026). More recent methods refine this view by accounting for the fact that not all tokens contribute equally to factual correctness. CCP aggregates uncertainty at the level of atomic claims (Fadeeva et al., 2024). CSL reweights token contributions using contextual relevance (Lin et al., 2024), while Self-Certainty measures confidence from the sharpness of the predicted token distribution (Kang et al., 2025). As noted by Dinh and Niehues (2025), token-level methods are subject to surface-level variability because probability mass can be distributed across multiple equally valid continuations. To address this issue, they propose BoostedProb, which aggregates probabilities when multiple tokens are valid. A third line of work moves from tokens to *semantic-level* agreement across sampled generations (Farquhar et al., 2024; Manakul et al., 2023; Qiu and Miikkulainen, 2024; Duan et al., 2024; Vashurin et al., 2025b). Semantic Entropy and Semantic Density estimate uncertainty in a semantic space rather than over tokens, making them more robust to paraphrases and surface-level variability (Qiu and Miikkulainen, 2024; Farquhar et al., 2024), while SAR also incorporates semantic relevance (Duan et al., 2024). CoCoA further argues that confidence-based and consistency-based signals are complementary and unifies these signals in a minimum Bayes risk formulation (Vashurin et al., 2025b). These approaches are more robust to paraphrases, but they rely on post-hoc semantic comparison after decoding, often through auxiliary models. Most closely related to our work are *representation-level* approaches, which argue that uncertainty is already encoded in the model's hidden states. Chen et al. (2024) introduce EigenScore and measure dispersion in the sentence-embedding space by using the log-determinant as an approximation to differential entropy. Similarly, Wang et al. (2025) propose using the Shannon entropy of the normalized singular values. Attention patterns themselves have also been observed to reflect hallucinations (Sriramanan et al., 2024; Vazhentsev et al., 2025), and these signals can be used to detect hallucinations from a single forward pass. Sriramanan et al. (2024) propose AttentionScore, which applies the same entropy-based metric introduced in EigenScore to the model's self-attention kernel maps, while Vazhentsev et al. (2025) propose RAUQ, which computes uncertainty by aggregating signals from specific uncertainty-aware heads that show a drop in attention to preceding tokens. Subsequent work shows that internal representations can reveal hallucination risk before generation (Ji et al., 2024) and can even support latent-space steering for detection (Park et al., 2025). Further evidence that information relevant to hallucination detection is present in the model's internal representations (Zhang et al., 2025; Du et al., 2024) also comes from probe-based methods (Azaria and Mitchell, 2023; Su et al., 2024; Kossen et al., 2024). Our method follows this representation-level perspective, but replaces entropy or probe-based scoring with a coverage view: uncertainty is quantified by how broadly generations span the model's latent semantic space, connecting LLM uncertainty quantification to exploration in reinforcement learning.

3 Method

Overview. Continuous control and reinforcement learning have extensively dealt with the question of how to characterize uncertainty in the context of arbitrating between exploratory and exploitative behavior. Generally, higher uncertainty is associ-

ated with trajectories whose visited states span a broader region of representation space. We transfer this intuition to uncertainty quantification for large language models by replacing a temporal trajectory of visited states with a set of independently sampled generations for the same prompt. Semantic Coverage first maps each generation to an internal response-level representation by mean-pooling its token-level hidden states, and then uncertainty is quantified by how broadly these representations span the model's latent semantic space. When the model is confident, sampled generations should correspond to a stable semantic answer and their representations should remain close in the space. When the model is uncertain, sampled generations may reflect multiple plausible semantic answers, producing broader coverage in latent semantic space and reflecting greater semantic uncertainty.

3.1 Preliminaries

Let x be an input prompt, and let $p_\theta(\cdot \mid x)$ denote the conditional distribution over completions induced by a language model with parameters θ. We consider N independent generations $(y^{(1)}, \ldots, y^{(N)})$ sampled under the fixed sampling policy $p_\theta(\cdot \mid x)$. Let $h_{l,t}(x, y) \in \mathbb{R}^d$ be the hidden representation of the LLM at layer l and position t for generation y conditioned on prompt x. To map each generated sequence to a single response-level semantic state, Semantic Coverage first mean-pools the hidden representations across generated tokens, following prior work on sequence-level embeddings (Ren et al., 2022). This mean-pooling step is the first component of Semantic Coverage and provides a simple fixed-dimensional summary that aggregates information across the full response trajectory on which the second spectral-summary component is then computed.

$$h_l(x, y) := \frac{1}{T(y)} \sum_{t=1}^{T(y)} h_{l,t}(x, y) \in \mathbb{R}^d,$$

where $T(y)$ is the number of generated tokens in the response y, and $h_l(x, y)$ is a trajectory-level semantic representation of the response y at layer l that aggregates information over the entire generated response. Given N sampled generations, we collect their semantic representations into the matrix

$$H_l(x) = \begin{bmatrix} h_l(x, y^{(1)})^\top \\ \vdots \\ h_l(x, y^{(N)})^\top \end{bmatrix} \in \mathbb{R}^{N \times d}.$$

We then center the representation matrix and define the empirical covariance matrix of the semantic representations at layer l as

$$\hat{\Sigma}_l(x) = \frac{1}{N} \tilde{H}_l(x)^\top \tilde{H}_l(x) \in \mathbb{R}^{d \times d},$$

where $\tilde{H}_l(x)$ is the centered embedding matrix.

3.2 Semantic Coverage

Following Wagenmaker et al. (2025), we define the coverage of the LLM's latent semantic space as the projection of the covariance matrix of one group of generations onto the semantic representation space defined by a second group of generations:

$$\mathrm{cov}_{l,\gamma}(x) := \frac{1}{\mathrm{tr}\left(\left(U_\epsilon^\top \hat{\Sigma}'_l(x) U_\epsilon + \gamma I\right)^{-1}\right)}, \quad (1)$$

where $\hat{\Sigma}_l(x)$ and $\hat{\Sigma}'_l(x)$ denote the covariances of a first and second group of generations, respectively, while U_ϵ denotes the singular vectors of $\hat{\Sigma}_l(x)$, whose corresponding singular values are at least ϵ. As the number of sampled generations grows, the two empirical covariance estimates $\hat{\Sigma}_l(x)$ and $\hat{\Sigma}'_l(x)$ converge to the same population covariance $\Sigma_l(x)$. Therefore, the projected quantity $U_\epsilon^\top \hat{\Sigma}'_l(x) U_\epsilon$ approaches the spectrum of $\Sigma_l(x)$ as $\epsilon \to 0$. This motivates defining the spectral-summary component of Semantic Coverage directly as a function of the spectrum of the empirical covariance from a single group of generations, avoiding the separate projection step while preserving the same asymptotic behavior:

$$\mathrm{cov}_{l,\gamma}(x) := \frac{1}{\mathrm{tr}\left(\left(\hat{\Sigma}_l(x) + \gamma I\right)^{-1}\right)}. \quad (2)$$

Here, $\gamma > 0$ is a regularization constant introduced to ensure numerical stability, especially in the common regime $N \ll d$, where $\hat{\Sigma}_l(x)$ is rank-deficient. Since

$$\mathrm{cov}_{l,\gamma}(x) = \left(\sum_{i=1}^{d} \frac{1}{\lambda_i + \gamma}\right)^{-1},$$

adding γ reduces the sensitivity of the metric to very small eigenvalues. In particular, for $\lambda_i \ll \gamma$, the contribution $1/(\lambda_i + \gamma)$ is approximately $1/\gamma$, reducing sensitivity to differences among very small eigenvalues, acting as a cutoff that determines when a direction counts as informative. For fixed

total variance, Semantic Coverage increases as variance is distributed more evenly across eigendirections. Further intuition and a latent-mixture interpretation are provided in Appendix A.

Volume vs. Shape. Despite their different motivations and derivations, both Semantic Coverage and EigenScore (Chen et al., 2024) make use of spectral information from the empirical covariance. EigenScore uses the log-determinant, $\frac{1}{d}\sum_i \log \lambda_i$, which corresponds to the geometric mean of the eigenvalues and to the log-volume of the covariance ellipsoid; Semantic Coverage uses a quantity proportional to the regularized harmonic mean, $(\sum_i 1/(\lambda_i))^{-1}$, placing greater relative emphasis on smaller eigenvalues (Appendix E.1).

4 Experiments

4.1 Experimental Setup

Models. We use three open-weights models from the Qwen (Yang et al., 2025) and Granite (Mishra et al., 2024) families, with 4B, 8B, and 14B parameters. In particular, we consider the pre-trained versions of `Qwen3-4B` and `Qwen3-14B` and the instruction-tuned version of `Granite3.3-8B`.

Datasets. Following Chen et al. (2024); Kossen et al. (2024); Xue et al. (2025); Duan et al. (2024); Qiu and Miikkulainen (2024), we focus our analysis on widely used Question-Answering (QA) datasets. In particular, we employ the test split of CoQA (Reddy et al., 2018), a conversational QA dataset that emphasizes multi-turn dialogue; SimpleQA (Wei et al., 2024), which emphasizes factuality; Natural Questions (Kwiatkowski et al., 2019), which emphasizes realistic search-style QA; and TriviaQA (Joshi et al., 2017), which emphasizes trivia knowledge. Following Vashurin et al. (2025a), we limit the evaluation set of each dataset to 2,000 examples and report the standard deviation across five runs.

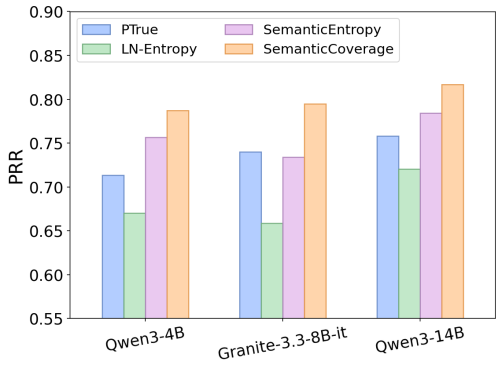
Baselines. We compare our method against a diverse set of baselines spanning the main families of uncertainty quantification methods for LLMs. From the *self-evaluation* family, we include Pr(True) (Kadavath et al., 2022). From the *token-level* family, we consider Perplexity (Ren et al., 2022), Length-Normalized Entropy (Malinin and Gales, 2020), G-NLL (Aichberger et al., 2026), CSL (Lin et al., 2024), and BoostedProb (Dinh and Niehues, 2025). From the *semantic-level* family, we include Semantic Entropy (Farquhar et al., 2024), Semantic Density (Qiu and Miikkulainen, 2024), SentenceSAR (Duan et al., 2024), and CoCoA using Perplexity as the confidence signal (Vashurin et al., 2025b). Finally, from the *representation-level* family, we consider EigenScore (Chen et al., 2024) and AttentionScore (Sriramanan et al., 2024), which derive uncertainty estimates directly from internal model representations.

Evaluation Metrics. Following Vashurin et al. (2025a,b), we employ Prediction Rejection Ratio (PRR) (Malinin and Gales, 2020). PRR measures how well uncertainty scores identify high-quality predictions. It is based on the prediction-rejection curve, which tracks the average prediction quality as increasingly uncertain samples are rejected. PRR normalizes the area under this curve relative to a random baseline and an oracle ranking induced by the true quality scores.

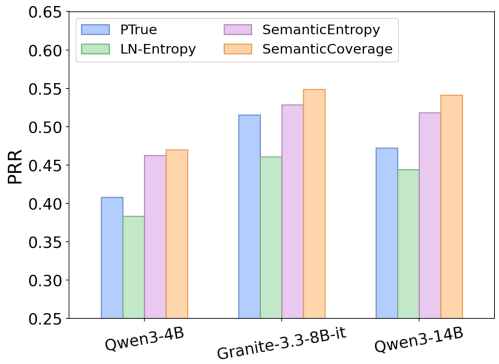
$$\text{PRR} = \frac{\text{AUC}_{\text{uncertainty}} - \text{AUC}_{\text{random}}}{\text{AUC}_{\text{oracle}} - \text{AUC}_{\text{random}}}$$

Quality Measures. Following Vashurin et al. (2025a); Fadeeva et al. (2023), we evaluate generation quality using BERTScore (Zhang et al., 2020) and AlignScore (Zha et al., 2023). BERTScore evaluates generated text by comparing candidate and reference tokens using contextual embeddings, and it has been shown to be more robust than traditional overlap-based metrics in paraphrastic settings (Zhang et al., 2020). AlignScore measures how well the generated answer is supported by the reference, providing a complementary signal that is more sensitive to factual consistency and has a higher correlation with human annotators (Zha et al., 2023). We use these two measures as continuous quality scores for each prediction.

Implementation Details. We generate answers using greedy decoding. For sampling-based methods, we draw 10 generations per example with temperature set to 1.0 and top-k set to 50. To ensure a fair comparison, all uncertainty quantification methods are evaluated on the same generated outputs. Unless otherwise noted, all remaining baseline hyperparameters are kept at the default values of their respective official implementations. Following Chen et al. (2024), we extract sentence embeddings from the middle layer of each model. For efficiency, we compute the covariance spectrum in



(a) Avg PRR with BERTScore



(b) Avg PRR with AlignScore

Figure 2: Average PRR across datasets for the three evaluated models (`Qwen3-4B`, `Granite-3.3-8B-it`, and `Qwen3-14B`) for four uncertainty quantification methods: Pr(True), LN-Entropy, Semantic Entropy, and Semantic Coverage. Semantic Coverage achieves the strongest overall performance. Top: Average PRR with BERTScore. Bottom: Average PRR with AlignScore.

sample space using the corresponding Gram matrix. For our method, we set the regularization parameter to 1.0, a sensitivity analysis is provided in Appendix C.

4.2 Results

Tables 1–3 and Figure 2 report the main results. Overall, Semantic Coverage is the strongest and most consistent uncertainty estimator across models, datasets, and evaluation metrics. Across the 24 reported settings, it achieves the best result in 17 of them and shows a consistent improvement over the strongest representation-level baseline, EigenScore. On `Qwen3-14B`, Semantic Coverage obtains the best PRR in 6 out of 8 settings. The same pattern holds for the smaller `Qwen3-4B`, where Semantic Coverage is best in 6 out of 8 PRR settings. Importantly, this trend also holds for the `Granite-3.3-8B-it` model, where Semantic Coverage remains the top method in 5 out of 8 settings. A pattern across method families also emerges from the results. Self-evaluation and token-level methods, such as Pr(True), Perplexity, Length-Normalized Entropy, CSL, and BoostedProb, are generally less robust, especially on Natural Questions and SimpleQA for the `Qwen3` models. Semantic-level methods, in particular Semantic Entropy and Semantic Density, provide strong baselines. Semantic Entropy tends to perform better than Semantic Coverage on the Natural Questions benchmark for both the 4B and 14B `Qwen3` models when using AlignScore as the quality metric, and also on SimpleQA for the 4B and 8B models. Notably, however, these methods still underperform Semantic Coverage when BERTScore is used as the quality metric. The best overall performance is typically obtained by *representation-level* methods. Within this family, Semantic Coverage consistently outperforms both EigenScore and AttentionScore, indicating the effectiveness of our method.

4.3 Ablation Studies

Unless otherwise stated, all ablation studies are conducted on a held-out subset of CoQA using BERTScore as the quality signal and PRR as the evaluation metric.

Trajectory View. So far, we have considered self-consistency at the level of multiple independent generations from the same prompt. However, Semantic Coverage can also be computed by treating a single LLM generation as a temporal trajectory. The same intuition that motivates the use of coverage across multiple generations also applies to this trajectory-based view of a single generation. In particular, if we view the sequence of internal representations produced during generation as a trajectory, we may expect an uncertain model to span a broader region of its latent semantic space. To test this idea, we compute Semantic Coverage using the internal representations of the tokens in a single generation. We consider two variants: one based on the last 10 tokens, where positions are filled from the context if the answer is shorter than 10 tokens; and one based on all generated tokens. Table 4 shows that both trajectory-level variants of Semantic Coverage consistently outperform methods such as Perplexity, LN-Entropy, and AttentionScore. For BoostedProb, Pr(True) and G-NLL, the results are slightly more nuanced, with these methods alternating with Semantic Coverage in the ranking. In particular G-NLL has strong per-

PRR

	CoQA		Natural Questions		SimpleQA		TriviaQA		Mean
Method	BERTScore	AlignScore	BERTScore	AlignScore	BERTScore	AlignScore	BERTScore	AlignScore	
Pr(True)	$0.847_{\pm 0.005}$	$0.815_{\pm 0.008}$	$0.641_{\pm 0.018}$	$0.219_{\pm 0.009}$	$0.677_{\pm 0.007}$	$0.085_{\pm 0.003}$	$0.867_{\pm 0.009}$	$0.770_{\pm 0.009}$	0.615
Perplexity	$0.775_{\pm 0.010}$	$0.813_{\pm 0.007}$	$0.628_{\pm 0.018}$	$0.238_{\pm 0.011}$	$0.597_{\pm 0.006}$	$0.077_{\pm 0.006}$	$0.811_{\pm 0.011}$	$0.693_{\pm 0.009}$	0.579
LN-Entropy	$0.776_{\pm 0.010}$	$0.795_{\pm 0.010}$	$0.651_{\pm 0.015}$	$0.242_{\pm 0.008}$	$0.654_{\pm 0.008}$	$0.076_{\pm 0.005}$	$0.801_{\pm 0.011}$	$0.663_{\pm 0.011}$	0.582
G-NLL	$0.814_{\pm 0.009}$	$0.827_{\pm 0.006}$	$0.663_{\pm 0.016}$	$0.272_{\pm 0.011}$	$0.640_{\pm 0.005}$	$0.089_{\pm 0.008}$	$0.834_{\pm 0.010}$	$0.726_{\pm 0.011}$	0.608
CSL	$0.782_{\pm 0.011}$	$0.807_{\pm 0.007}$	$0.629_{\pm 0.018}$	$0.236_{\pm 0.012}$	$0.612_{\pm 0.007}$	$0.075_{\pm 0.004}$	$0.810_{\pm 0.012}$	$0.689_{\pm 0.011}$	0.580
BoostedProb	$0.769_{\pm 0.010}$	$0.817_{\pm 0.005}$	$0.641_{\pm 0.016}$	$0.243_{\pm 0.008}$	$0.666_{\pm 0.007}$	$0.079_{\pm 0.004}$	$0.833_{\pm 0.010}$	$0.729_{\pm 0.008}$	0.597
SemanticEntropy	$0.838_{\pm 0.006}$	$0.837_{\pm 0.007}$	$0.732_{\pm 0.010}$	$0.353_{\pm 0.005}$	$0.703_{\pm 0.009}$	$0.112_{\pm 0.009}$	$0.864_{\pm 0.006}$	$0.770_{\pm 0.008}$	0.651
SemanticDensity	$0.856_{\pm 0.008}$	$0.848_{\pm 0.009}$	$0.701_{\pm 0.013}$	$\mathbf{0.357}_{\pm 0.010}$	$0.612_{\pm 0.007}$	$0.108_{\pm 0.006}$	$0.884_{\pm 0.005}$	$0.810_{\pm 0.004}$	0.647
SentenceSAR	$0.820_{\pm 0.008}$	$0.827_{\pm 0.009}$	$0.730_{\pm 0.010}$	$0.344_{\pm 0.005}$	$0.711_{\pm 0.007}$	$\mathbf{0.118}_{\pm 0.008}$	$0.850_{\pm 0.007}$	$0.748_{\pm 0.009}$	0.643
CoCoAPPL	$0.807_{\pm 0.008}$	$0.843_{\pm 0.005}$	$0.661_{\pm 0.015}$	$0.292_{\pm 0.010}$	$0.613_{\pm 0.006}$	$0.089_{\pm 0.006}$	$0.861_{\pm 0.008}$	$0.775_{\pm 0.009}$	0.618
AttnScore	$0.821_{\pm 0.010}$	$0.799_{\pm 0.009}$	$0.690_{\pm 0.015}$	$0.252_{\pm 0.009}$	$0.684_{\pm 0.009}$	$0.083_{\pm 0.008}$	$0.788_{\pm 0.013}$	$0.632_{\pm 0.012}$	0.594
EigenScore	$0.895_{\pm 0.006}$	$0.871_{\pm 0.006}$	$0.728_{\pm 0.012}$	$0.312_{\pm 0.010}$	$0.716_{\pm 0.006}$	$0.099_{\pm 0.004}$	$0.895_{\pm 0.005}$	$0.815_{\pm 0.004}$	0.666
SemanticCoverage	$\mathbf{0.902}_{\pm 0.005}$	$\mathbf{0.881}_{\pm 0.006}$	$\mathbf{0.739}_{\pm 0.011}$	$0.341_{\pm 0.010}$	$\mathbf{0.725}_{\pm 0.007}$	$0.116_{\pm 0.004}$	$\mathbf{0.901}_{\pm 0.005}$	$\mathbf{0.826}_{\pm 0.004}$	**0.679**

Table 1: Performance of uncertainty quantification methods using the PRR evaluation metric for `Qwen3-14B` across QA benchmarks. Each entry shows mean with standard deviation; the final column reports the average across datasets and quality measures.

PRR

	CoQA		Natural Questions		SimpleQA		TriviaQA		Mean
Method	BERTScore	AlignScore	BERTScore	AlignScore	BERTScore	AlignScore	BERTScore	AlignScore	
Pr(True)	$0.838_{\pm 0.004}$	$0.787_{\pm 0.008}$	$0.614_{\pm 0.004}$	$0.192_{\pm 0.011}$	$0.621_{\pm 0.034}$	$0.066_{\pm 0.005}$	$0.780_{\pm 0.022}$	$0.587_{\pm 0.007}$	0.561
Perplexity	$0.696_{\pm 0.008}$	$0.784_{\pm 0.010}$	$0.553_{\pm 0.003}$	$0.194_{\pm 0.008}$	$0.536_{\pm 0.045}$	$0.062_{\pm 0.003}$	$0.712_{\pm 0.022}$	$0.508_{\pm 0.011}$	0.506
LN-Entropy	$0.731_{\pm 0.005}$	$0.780_{\pm 0.009}$	$0.609_{\pm 0.004}$	$0.196_{\pm 0.012}$	$0.620_{\pm 0.037}$	$0.056_{\pm 0.003}$	$0.721_{\pm 0.021}$	$0.502_{\pm 0.007}$	0.527
G-NLL	$0.762_{\pm 0.006}$	$0.813_{\pm 0.012}$	$0.604_{\pm 0.004}$	$0.225_{\pm 0.011}$	$0.596_{\pm 0.040}$	$0.073_{\pm 0.004}$	$0.765_{\pm 0.020}$	$0.593_{\pm 0.007}$	0.554
CSL	$0.714_{\pm 0.009}$	$0.779_{\pm 0.012}$	$0.578_{\pm 0.003}$	$0.203_{\pm 0.009}$	$0.575_{\pm 0.043}$	$0.061_{\pm 0.004}$	$0.718_{\pm 0.022}$	$0.514_{\pm 0.011}$	0.518
BoostedProb	$0.700_{\pm 0.009}$	$0.794_{\pm 0.009}$	$0.580_{\pm 0.005}$	$0.207_{\pm 0.009}$	$0.608_{\pm 0.035}$	$0.064_{\pm 0.004}$	$0.771_{\pm 0.017}$	$0.605_{\pm 0.010}$	0.541
SemanticEntropy	$0.833_{\pm 0.005}$	$0.841_{\pm 0.007}$	$0.684_{\pm 0.004}$	$0.260_{\pm 0.008}$	$0.693_{\pm 0.027}$	$\mathbf{0.083}_{\pm 0.004}$	$0.815_{\pm 0.014}$	$0.667_{\pm 0.006}$	0.609
SemanticDensity	$0.811_{\pm 0.011}$	$0.819_{\pm 0.008}$	$0.626_{\pm 0.006}$	$\mathbf{0.271}_{\pm 0.008}$	$0.555_{\pm 0.039}$	$0.077_{\pm 0.004}$	$0.816_{\pm 0.017}$	$0.675_{\pm 0.005}$	0.581
SentenceSAR	$0.797_{\pm 0.006}$	$0.822_{\pm 0.006}$	$0.664_{\pm 0.005}$	$0.242_{\pm 0.008}$	$0.667_{\pm 0.030}$	$0.080_{\pm 0.004}$	$0.767_{\pm 0.017}$	$0.585_{\pm 0.010}$	0.578
CoCoAPPL	$0.752_{\pm 0.006}$	$0.831_{\pm 0.009}$	$0.590_{\pm 0.003}$	$0.236_{\pm 0.010}$	$0.549_{\pm 0.043}$	$0.068_{\pm 0.003}$	$0.799_{\pm 0.016}$	$0.656_{\pm 0.006}$	0.560
AttnScore	$0.807_{\pm 0.006}$	$0.780_{\pm 0.008}$	$0.661_{\pm 0.004}$	$0.193_{\pm 0.012}$	$0.669_{\pm 0.029}$	$0.070_{\pm 0.006}$	$0.723_{\pm 0.021}$	$0.487_{\pm 0.008}$	0.549
EigenScore	$0.883_{\pm 0.005}$	$0.843_{\pm 0.009}$	$0.680_{\pm 0.003}$	$0.215_{\pm 0.009}$	$0.706_{\pm 0.027}$	$0.075_{\pm 0.004}$	$0.828_{\pm 0.017}$	$0.675_{\pm 0.006}$	0.613
SemanticCoverage	$\mathbf{0.894}_{\pm 0.005}$	$\mathbf{0.860}_{\pm 0.008}$	$\mathbf{0.691}_{\pm 0.004}$	$0.236_{\pm 0.011}$	$\mathbf{0.719}_{\pm 0.026}$	$0.077_{\pm 0.003}$	$\mathbf{0.845}_{\pm 0.014}$	$\mathbf{0.708}_{\pm 0.005}$	**0.629**

Table 2: Performance of uncertainty quantification methods using the PRR evaluation metric for `Qwen3-4B` across QA benchmarks. Each entry shows mean with standard deviation; the final column reports the average across datasets and quality measures.

PRR

	CoQA		Natural Questions		SimpleQA		TriviaQA		Mean
Method	BERTScore	AlignScore	BERTScore	AlignScore	BERTScore	AlignScore	BERTScore	AlignScore	
Pr(True)	$0.801_{\pm 0.006}$	$0.854_{\pm 0.007}$	$0.633_{\pm 0.018}$	$0.311_{\pm 0.008}$	$0.638_{\pm 0.046}$	$0.070_{\pm 0.007}$	$0.887_{\pm 0.006}$	$0.825_{\pm 0.008}$	0.627
Perplexity	$0.611_{\pm 0.011}$	$0.790_{\pm 0.008}$	$0.567_{\pm 0.021}$	$0.264_{\pm 0.013}$	$0.599_{\pm 0.051}$	$0.069_{\pm 0.004}$	$0.828_{\pm 0.007}$	$0.754_{\pm 0.005}$	0.560
LN-Entropy	$0.631_{\pm 0.010}$	$0.782_{\pm 0.010}$	$0.563_{\pm 0.020}$	$0.252_{\pm 0.009}$	$0.619_{\pm 0.049}$	$0.069_{\pm 0.006}$	$0.821_{\pm 0.007}$	$0.739_{\pm 0.005}$	0.559
G-NLL	$0.676_{\pm 0.012}$	$0.812_{\pm 0.009}$	$0.599_{\pm 0.020}$	$0.291_{\pm 0.014}$	$0.627_{\pm 0.047}$	$0.079_{\pm 0.003}$	$0.848_{\pm 0.007}$	$0.779_{\pm 0.007}$	0.589
CSL	$0.611_{\pm 0.013}$	$0.788_{\pm 0.007}$	$0.568_{\pm 0.022}$	$0.264_{\pm 0.012}$	$0.617_{\pm 0.050}$	$0.067_{\pm 0.005}$	$0.834_{\pm 0.008}$	$0.760_{\pm 0.009}$	0.564
BoostedProb	$0.605_{\pm 0.012}$	$0.787_{\pm 0.009}$	$0.569_{\pm 0.020}$	$0.276_{\pm 0.012}$	$0.644_{\pm 0.044}$	$0.066_{\pm 0.002}$	$0.839_{\pm 0.008}$	$0.774_{\pm 0.005}$	0.570
SemanticEntropy	$0.726_{\pm 0.012}$	$0.841_{\pm 0.007}$	$0.664_{\pm 0.015}$	$0.360_{\pm 0.007}$	$0.674_{\pm 0.042}$	$\mathbf{0.101}_{\pm 0.010}$	$0.872_{\pm 0.005}$	$0.812_{\pm 0.005}$	0.631
SemanticDensity	$0.739_{\pm 0.011}$	$0.846_{\pm 0.005}$	$0.674_{\pm 0.016}$	$\mathbf{0.388}_{\pm 0.013}$	$0.617_{\pm 0.048}$	$0.096_{\pm 0.008}$	$0.909_{\pm 0.007}$	$\mathbf{0.858}_{\pm 0.011}$	0.641
SentenceSAR	$0.681_{\pm 0.011}$	$0.819_{\pm 0.007}$	$0.631_{\pm 0.017}$	$0.326_{\pm 0.009}$	$0.656_{\pm 0.044}$	$0.094_{\pm 0.007}$	$0.858_{\pm 0.005}$	$0.792_{\pm 0.006}$	0.607
CoCoAPPL	$0.662_{\pm 0.012}$	$0.820_{\pm 0.008}$	$0.608_{\pm 0.019}$	$0.320_{\pm 0.014}$	$0.617_{\pm 0.049}$	$0.085_{\pm 0.004}$	$0.873_{\pm 0.003}$	$0.823_{\pm 0.003}$	0.601
AttnScore	$0.715_{\pm 0.014}$	$0.813_{\pm 0.004}$	$0.687_{\pm 0.013}$	$0.295_{\pm 0.012}$	$0.675_{\pm 0.040}$	$0.071_{\pm 0.004}$	$0.839_{\pm 0.008}$	$0.744_{\pm 0.009}$	0.605
EigenScore	$0.822_{\pm 0.006}$	$0.872_{\pm 0.004}$	$0.710_{\pm 0.011}$	$0.363_{\pm 0.014}$	$0.709_{\pm 0.037}$	$0.085_{\pm 0.005}$	$0.903_{\pm 0.005}$	$0.842_{\pm 0.006}$	0.663
SemanticCoverage	$\mathbf{0.830}_{\pm 0.006}$	$\mathbf{0.879}_{\pm 0.002}$	$\mathbf{0.716}_{\pm 0.011}$	$0.376_{\pm 0.015}$	$\mathbf{0.723}_{\pm 0.035}$	$0.087_{\pm 0.005}$	$\mathbf{0.910}_{\pm 0.005}$	$0.853_{\pm 0.005}$	**0.672**

Table 3: Performance of uncertainty quantification methods using the PRR evaluation metric for `Granite-3.3-8B-it` across QA benchmarks. Each entry shows mean with standard deviation; the final column reports the average across datasets and quality measures.

PRR

Method	CoQA		Natural Questions		SimpleQA		TriviaQA		Mean
	BERTScore	AlignScore	BERTScore	AlignScore	BERTScore	AlignScore	BERTScore	AlignScore	
Pr(True)	$0.838_{\pm 0.004}$	$0.787_{\pm 0.008}$	$0.614_{\pm 0.004}$	$0.192_{\pm 0.011}$	$0.621_{\pm 0.034}$	$0.066_{\pm 0.005}$	$\mathbf{0.780}_{\pm 0.022}$	$0.587_{\pm 0.007}$	0.561
Perplexity	$0.696_{\pm 0.008}$	$0.784_{\pm 0.010}$	$0.553_{\pm 0.003}$	$0.194_{\pm 0.008}$	$0.536_{\pm 0.045}$	$0.062_{\pm 0.003}$	$0.712_{\pm 0.022}$	$0.508_{\pm 0.011}$	0.506
LN-Entropy	$0.731_{\pm 0.005}$	$0.780_{\pm 0.009}$	$0.609_{\pm 0.004}$	$0.196_{\pm 0.012}$	$0.620_{\pm 0.037}$	$0.056_{\pm 0.003}$	$0.721_{\pm 0.021}$	$0.502_{\pm 0.007}$	0.527
G-NLL	$0.762_{\pm 0.006}$	$\mathbf{0.813}_{\pm 0.012}$	$0.604_{\pm 0.004}$	$\mathbf{0.225}_{\pm 0.011}$	$0.596_{\pm 0.040}$	$0.073_{\pm 0.004}$	$0.765_{\pm 0.020}$	$0.593_{\pm 0.007}$	0.554
CSL	$0.714_{\pm 0.009}$	$0.779_{\pm 0.012}$	$0.578_{\pm 0.003}$	$0.203_{\pm 0.009}$	$0.575_{\pm 0.043}$	$0.061_{\pm 0.004}$	$0.718_{\pm 0.022}$	$0.514_{\pm 0.011}$	0.518
BoostedProb	$0.700_{\pm 0.009}$	$0.794_{\pm 0.009}$	$0.580_{\pm 0.005}$	$0.207_{\pm 0.009}$	$0.608_{\pm 0.035}$	$0.064_{\pm 0.004}$	$0.771_{\pm 0.017}$	$\mathbf{0.605}_{\pm 0.010}$	0.541
AttnScore	$0.807_{\pm 0.006}$	$0.780_{\pm 0.008}$	$0.661_{\pm 0.004}$	$0.193_{\pm 0.012}$	$0.669_{\pm 0.029}$	$0.070_{\pm 0.006}$	$0.723_{\pm 0.021}$	$0.487_{\pm 0.008}$	0.549
SemanticCoverageTraj (w=10)	$0.852_{\pm 0.007}$	$0.792_{\pm 0.008}$	$\mathbf{0.675}_{\pm 0.005}$	$0.213_{\pm 0.009}$	$0.682_{\pm 0.030}$	$0.073_{\pm 0.004}$	$0.746_{\pm 0.024}$	$0.513_{\pm 0.003}$	**0.568**
SemanticCoverageTraj (all)	$\mathbf{0.854}_{\pm 0.007}$	$0.794_{\pm 0.010}$	$0.673_{\pm 0.005}$	$0.212_{\pm 0.008}$	$\mathbf{0.684}_{\pm 0.030}$	$\mathbf{0.074}_{\pm 0.004}$	$0.740_{\pm 0.024}$	$0.508_{\pm 0.003}$	0.567

Table 4: Performance of Pr(True), token-level methods, AttentionScore, and trajectory-level variants of Semantic Coverage for `Qwen3-4B` across QA benchmarks. SemanticCoverageTraj (all) refers to using all the generated tokens, SemanticCoverageTraj (w=10) refers to using a window of the last 10 tokens. Each entry shows mean and standard deviation; the final column reports the average across datasets and quality measures.

formances with AlignScore on CoQA and Natural Question, instead BoostedProb and Pr(True), while underperforming relative to Semantic Coverage on CoQA, Natural Questions, and SimpleQA, they obtain higher PRR on TriviaQA. Notably, however, both trajectory variants of Semantic Coverage achieve the highest mean score among the evaluated methods.

Sensitivity to Hyperparameters. Figure 4 shows that varying the number of generations has only a limited effect on performance. Increasing the sample budget from 5 to 20 leads to small gains in PRR and does not change the ranking of the methods. As shown in Figure 3, Semantic Coverage's performance improves as the temperature increases from 0.1 to 1.0, after which it declines again. This trend is consistent with the role of temperature in controlling generation diversity: values that are too small produce responses that are too concentrated to reveal semantic uncertainty, while values that are too large introduce excess variability that weakens the uncertainty signal. We therefore use a temperature of 1.0 in the main experiments. Notably, Semantic Coverage tends to outperform the compared methods across both sampling temperatures and generation budgets.

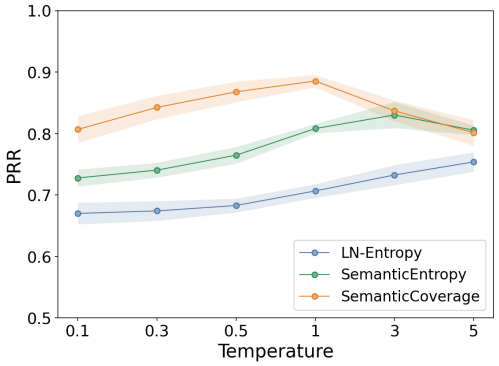


Figure 3: Sensitivity to temperature for `Qwen3-4B` on the CoQA dataset. Shaded regions indicate mean ± 1 std.

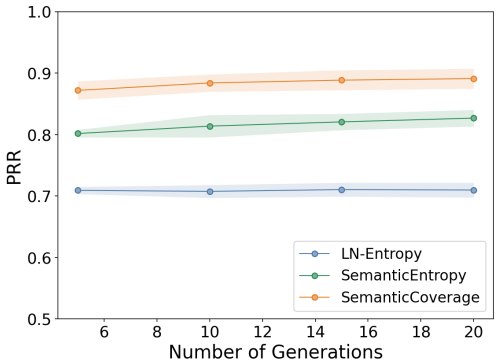


Figure 4: Sensitivity to number of generations for `Qwen3-4B` on the CoQA dataset. Shaded regions indicate mean ± 1 std.

5 Conclusion

We introduced Semantic Coverage, a white-box representation-level uncertainty estimator that measures how broadly multiple sampled generations span the model's latent semantic space. Across three open-weights LLMs and four short-form QA benchmarks, it consistently outperformed prior representation-level methods and achieved the strongest overall results among the evaluated methods. These findings suggest that uncertainty is reflected in how internal representations span latent semantic space, and that Semantic Coverage provides a strong signal for uncertainty quantification in LLMs.

Limitations

Our method relies on access to hidden representations to compute sentence-level embeddings and

estimate Semantic Coverage, and is therefore not directly applicable in black-box API settings. The method is also sampling-based, which requires multiple generations per prompt and therefore incurs additional inference cost with respect to methods that do not require multiple samples. We study uncertainty quantification in short-form question answering, following the standard evaluation setting used in prior work (Chen et al., 2024; Kossen et al., 2024; Xue et al., 2025; Duan et al., 2024; Qiu and Miikkulainen, 2024). While this enables controlled comparison against established baselines, real-world deployment often involves more complex settings. Finally, our evaluation relies on quality metrics. Although BERTScore and AlignScore are widely used, they do not capture every nuance of generation quality. Future work could address these limitations by developing black-box approximations of Semantic Coverage, extending the method to more complex settings, and incorporating additional metrics to provide a more comprehensive, risk-aware assessment of generation quality (Nitsure et al., 2024).

Ethical Considerations

We use publicly available open-weight models and established QA benchmarks, and our experiments are not designed to elicit harmful content. Nevertheless, LLM outputs can be incorrect, misleading, or harmful in downstream use. Uncertainty quantification methods are intended to support reliability assessment by flagging potentially uncertain generations, but they may produce false positives or false negatives and should not be treated as a guarantee of correctness.

Acknowledgments

This work was supported by the Swiss State Secretariat for Education, Research and Innovation (SERI) under SBFI No. 25.00154 and by the European Union's Horizon Europe Marie Skłodowska-Curie Actions (MSCA) through the ARMADA project under grant agreement No. 101168951.

References

Lukas Aichberger, Kajetan Schweighofer, and Sepp Hochreiter. 2026. Rethinking Uncertainty Estimation in LLMs: A Principled Single-Sequence Measure. *arXiv preprint*. ArXiv:2412.15176 [cs.LG] version: 3.

Sihem Amer-Yahia, Jasmina Bogojeska, Roberta Facchinetti, Valeria Franceschi, Aristides Gionis, Katja Hose, Georgia Koutrika, Roger Kouyos, Matteo Lissandrini, Silviu Maniu, Katsiaryna Mirylenka, Davide Mottin, Themis Palpanas, Mattia Rigotti, and Yannis Velegrakis. 2025. Towards Reliable Conversational Data Analytics.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. 2002. Finite-time Analysis of the Multiarmed Bandit Problem. *Machine Learning*, 47(2):235–256.

Amos Azaria and Tom Mitchell. 2023. The Internal State of an LLM Knows When It's Lying. *arXiv preprint*. ArXiv:2304.13734 [cs].

Charles W. Champ and Andrew V. Sills. 2024. The Generalized Laws of Total Variance and Total Covariance. In *Applied Mathematical Analysis and Computations I*, pages 243–254, Cham. Springer Nature Switzerland.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024. INSIDE: LLMs' Internal States Retain the Power of Hallucination Detection.

Tu Anh Dinh and Jan Niehues. 2025. Are Generative Models Underconfident? Better Quality Estimation with Boosted Model Probability. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3364–3382, Suzhou, China. Association for Computational Linguistics.

Xuefeng Du, Chaowei Xiao, and Yixuan Li. 2024. HaloScope: Harnessing Unlabeled LLM Generations for Hallucination Detection. *arXiv preprint*. ArXiv:2409.17504 [cs].

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2024. Shifting Attention to Relevance: Towards the Predictive Uncertainty Quantification of Free-Form Large Language Models. *arXiv preprint*. ArXiv:2307.01379 [cs].

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, and Maxim Panov. 2024. Fact-Checking the Output of Large Language Models via Token-Level Uncertainty Quantification. *arXiv preprint*. ArXiv:2403.04696 [cs].

Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, Timothy Baldwin, and Artem Shelmanov. 2023. LM-Polygraph: Uncertainty Estimation for Language Models. *arXiv preprint*. ArXiv:2311.07383 [cs].

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. Llama 3 Herd of Models. *arXiv preprint*. ArXiv:2407.21783 [cs].

Kai He, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, and Erik Cambria. 2025. A Survey of Large Language Models for Healthcare: from Data, Technology, and Applications to Accountability and Ethics. *arXiv preprint*. ArXiv:2310.05694 [cs].

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. *arXiv preprint*. ArXiv:2009.03300 [cs].

Ziwei Ji, Delong Chen, Etsuko Ishii, Samuel Cahyawijaya, Yejin Bang, Bryan Wilie, and Pascale Fung. 2024. LLM Internal States Reveal Hallucination Risk Faced With a Query.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language Models (Mostly) Know What They Know. *arXiv preprint*. ArXiv:2207.05221 [cs].

Zhewei Kang, Xuandong Zhao, and Dawn Song. 2025. Scalable Best-of-N Selection for Large Language Models via Self-Certainty. *arXiv preprint*. ArXiv:2502.18581 [cs].

Sanyam Kapoor, Nate Gruver, Manley Roberts, Katherine Collins, Arka Pal, Umang Bhatt, Adrian Weller, Samuel Dooley, Micah Goldblum, and Andrew Gordon Wilson. 2025. Large Language Models Must Be Taught to Know What They Don't Know. *arXiv preprint*. ArXiv:2406.08391 [cs].

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. 2024. Semantic Entropy Probes: Robust and Cheap Hallucination Detection in LLMs. *arXiv preprint*. ArXiv:2406.15927 [cs].

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Jinqi Lai, Wensheng Gan, Jiayang Wu, Zhenlian Qi, and Philip S. Yu. 2023. Large Language Models in Law: A Survey. *arXiv preprint*. ArXiv:2312.03718 [cs].

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024. Contextualized Sequence Likelihood: Enhanced Confidence Scores for Natural Language Generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10351–10368, Miami, Florida, USA. Association for Computational Linguistics.

Zhu Liu, Cunliang Kong, Ying Liu, and Maosong Sun. 2024. Fantastic Semantics and Where to Find Them: Investigating Which Layers of Generative LLMs Reflect Lexical Semantics. *arXiv preprint*. ArXiv:2403.01509 [cs].

Andrey Malinin and Mark Gales. 2020. Uncertainty Estimation in Autoregressive Structured Prediction.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models.

Mayank Mishra, Matt Stallone, Gaoyuan Zhang, Yikang Shen, Aditya Prasad, Adriana Meza Soria, Michele Merler, Parameswaran Selvam, Saptha Surendran, Shivdeep Singh, Manish Sethi, Xuan-Hong Dang, Pengyuan Li, Kun-Lung Wu, Syed Zawad, Andrew Coleman, Matthew White, Mark Lewis, Raju Pavuluri, and 27 others. 2024. Granite Code Models: A Family of Open Foundation Models for Code Intelligence. *arXiv preprint*. ArXiv:2405.04324 [cs].

Yuqi Nie, Yaxuan Kong, Xiaowen Dong, John M. Mulvey, H. Vincent Poor, Qingsong Wen, and Stefan Zohren. 2024. A Survey of Large Language Models for Financial Applications: Progress, Prospects and Challenges. *arXiv preprint*. ArXiv:2406.11903 [q-fin].

Apoorva Nitsure, Youssef Mroueh, Mattia Rigotti, Kristjan Greenewald, Brian Belgodere, Mikhail Yurochkin, Jiri Navratil, Igor Melnyk, and Jerret Ross. 2024. Risk Aware Benchmarking of Large Language Models. *arXiv preprint*. ArXiv:2310.07132 [cs].

Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. 2016. Deep Exploration via Bootstrapped DQN. *arXiv preprint*. ArXiv:1602.04621 [cs].

Seongheon Park, Xuefeng Du, Min-Hsuan Yeh, Haobo Wang, and Yixuan Li. 2025. Steer LLM Latents for Hallucination Detection. *arXiv preprint*. ArXiv:2503.01917 [cs].

Xin Qiu and Risto Miikkulainen. 2024. Semantic Density: Uncertainty Quantification for Large Language Models through Confidence Measurement in Semantic Space. *arXiv preprint*. ArXiv:2405.13845 [cs].

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2018. CoQA: A Conversational Question Answering Challenge.

Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J. Liu. 2022. Out-of-Distribution Detection and Selective Generation for Conditional Language Models.

Jie Ren, Yao Zhao, Tu Vu, Peter J. Liu, and Balaji Lakshminarayanan. 2023. Self-Evaluation Improves Selective Generation in Large Language Models. *arXiv preprint*. ArXiv:2312.09300 [cs].

Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z. Ren, and Anirudha Majumdar. 2025. A Survey on Uncertainty Quantification of Large Language Models: Taxonomy, Open Research Challenges, and Future Directions. *arXiv preprint*. ArXiv:2412.05563 [cs].

Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. 2024. LLM-Check: Investigating Detection of Hallucinations in Large Language Models. *Advances in Neural Information Processing Systems*, 37:34188–34216.

Weihang Su, Changyue Wang, Qingyao Ai, Yiran HU, Zhijing Wu, Yujia Zhou, and Yiqun Liu. 2024. Unsupervised Real-Time Hallucination Detection based on the Internal States of Large Language Models. *arXiv preprint*. ArXiv:2403.06448 [cs].

Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Lyudmila Rvanova, Akim Tsvigun, Daniil Vasilev, Rui Xing, Abdelrahman Boda Sadallah, Kirill Grishchenkov, Sergey Petrakov, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, Maxim Panov, and Artem Shelmanov. 2025a. Benchmarking Uncertainty Quantification Methods for Large Language Models with LM-Polygraph. *Transactions of the Association for Computational Linguistics*, 13:220–248. ArXiv:2406.15627 [cs].

Roman Vashurin, Maiya Goloburda, Albina Ilina, Aleksandr Rubashevskii, Preslav Nakov, Artem Shelmanov, and Maxim Panov. 2025b. Uncertainty Quantification for LLMs through Minimum Bayes Risk: Bridging Confidence and Consistency.

Artem Vazhentsev, Lyudmila Rvanova, Gleb Kuzmin, Ekaterina Fadeeva, Ivan Lazichny, Alexander Panchenko, Maxim Panov, Timothy Baldwin, Mrinmaya Sachan, Preslav Nakov, and Artem Shelmanov. 2025. Uncertainty-Aware Attention Heads: Efficient Unsupervised Uncertainty Quantification for LLMs.

Andrew Wagenmaker, Zhiyuan Zhou, and Sergey Levine. 2025. Behavioral Exploration: Learning to Explore via In-Context Adaptation. *arXiv preprint*. ArXiv:2507.09041 [cs].

Rui Wang, Zeming Wei, Guanzhang Yue, and Meng Sun. 2025. Revisiting Hallucination Detection with Effective Rank-based Uncertainty. *arXiv preprint*. ArXiv:2510.08389 [cs].

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024. Measuring short-form factuality in large language models.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. *arXiv preprint*. ArXiv:2306.13063 [cs].

Yihao Xue, Kristjan Greenewald, Youssef Mroueh, and Baharan Mirzasoleiman. 2025. Verify when Uncertain: Beyond Self-Consistency in Black Box Hallucination Detection. *arXiv preprint*. ArXiv:2502.15845 [cs].

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 Technical Report. *arXiv preprint*. ArXiv:2505.09388 [cs].

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating Factual Consistency with a Unified Alignment Function. *arXiv preprint*. ArXiv:2305.16739 [cs].

Fujie Zhang, Peiqi Yu, Biao Yi, Baolei Zhang, Tong Li, and Zheli Liu. 2025. Prompt-Guided Internal States for Hallucination Detection of Large Language Models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21806–21818, Vienna, Austria. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. *arXiv preprint*. ArXiv:1904.09675 [cs].

Rong Zhu and Mattia Rigotti. 2021. Deep Bandits Show-Off: Simple and Efficient Exploration with Deep Networks. *arXiv preprint*. ArXiv:2105.04683 [cs].

A Latent Semantic Mixture Decomposition

To build intuition for what Semantic Coverage captures, we consider a latent-variable model for sampled responses. Let $Z \in \{1, \dots, K\}$ denote a latent semantic mode, and let h be the representation of a sampled answer. Assume

$$h = \mu_Z + \varepsilon, \qquad \mathbb{E}[\varepsilon \mid Z] = 0,$$

where $\mu_Z := \mathbb{E}[h \mid Z]$ is the mode-conditional mean representation, and ε captures residual variation within a mode, such as phrasing, style, and sampling noise. Under these assumptions one can show that the following decomposition holds:

$$\mathrm{Cov}(h) = \underbrace{\mathrm{Cov}(\mu_Z)}_{\text{between-mode}} + \underbrace{\mathbb{E}[\mathrm{Cov}(\varepsilon \mid Z)]}_{\text{within-mode}}$$

(see Derivation in Appendix A.1). The first term captures how strongly distinct semantic modes are separated in representation space, while the second captures local variation within each mode. This decomposition helps interpret Semantic Coverage. If repeated generations correspond to paraphrases of essentially the same answer, then the mode means μ_Z remain close, and the between-mode term is small; variability is dominated by within-mode fluctuations. By contrast, when the model assigns substantial probability mass to several semantically distinct answers, the means μ_Z become more dispersed, increasing the between-mode covariance. This spreads the sampled representations across a broader set of directions in latent space. Since Semantic Coverage increases when variance is distributed across multiple directions rather than concentrated in a narrow subspace, it is expected to be sensitive to uncertainty arising from multiple plausible semantic modes. However, Semantic Coverage does not isolate semantic variation perfectly: large within-mode variability can also increase dispersion. Its usefulness therefore depends on the hypothesis that hidden-state representations separate genuinely distinct answers more strongly than they separate superficial variation, such as phrasing or style.

A.1 Derivation of Latent Semantic Mixture Decomposition

Let

$$\begin{aligned} h &= \mu_Z + \varepsilon, \\ \mu_Z &:= \mathbb{E}[h \mid Z], \\ \varepsilon &:= h - \mu_Z. \end{aligned}$$

Then

$$\begin{aligned} \mathbb{E}[\varepsilon \mid Z] &= \mathbb{E}[h - \mu_Z \mid Z] \\ &= \mathbb{E}[h \mid Z] - \mu_Z \\ &= 0. \end{aligned}$$

Using $h = \mu_Z + \varepsilon$, we have

$$\begin{aligned} \mathrm{Cov}(h) &= \mathrm{Cov}(\mu_Z + \varepsilon) \\ &= \mathrm{Cov}(\mu_Z) + \mathrm{Cov}(\varepsilon) \\ &\quad + \mathrm{Cov}(\mu_Z, \varepsilon) + \mathrm{Cov}(\varepsilon, \mu_Z). \end{aligned}$$

The cross-covariance terms vanish. Since μ_Z is a function of Z and $\mathbb{E}[\varepsilon \mid Z] = 0$,

$$\begin{aligned} \mathrm{Cov}(\mu_Z, \varepsilon) &= \mathbb{E}\Big[(\mu_Z - \mathbb{E}[\mu_Z])\varepsilon^\top\Big] \\ &= \mathbb{E}\Big[\mathbb{E}\Big[(\mu_Z - \mathbb{E}[\mu_Z])\varepsilon^\top \mid Z\Big]\Big] \\ &= \mathbb{E}\Big[(\mu_Z - \mathbb{E}[\mu_Z])\mathbb{E}[\varepsilon^\top \mid Z]\Big] \\ &= 0. \end{aligned}$$

Similarly, $\mathrm{Cov}(\varepsilon, \mu_Z) = 0$. Therefore,

$$\mathrm{Cov}(h) = \mathrm{Cov}(\mu_Z) + \mathrm{Cov}(\varepsilon).$$

Finally, applying the law of total covariance (Champ and Sills, 2024) to ε:

$$\begin{aligned} \mathrm{Cov}(\varepsilon) &= \mathbb{E}[\mathrm{Cov}(\varepsilon \mid Z)] + \mathrm{Cov}(\mathbb{E}[\varepsilon \mid Z]) \\ &= \mathbb{E}[\mathrm{Cov}(\varepsilon \mid Z)]\,. \end{aligned}$$

B Projection-based Coverage

Figure 5 empirically shows that the projection-based coverage and Semantic Coverage converge as the number of generations becomes sufficiently large.

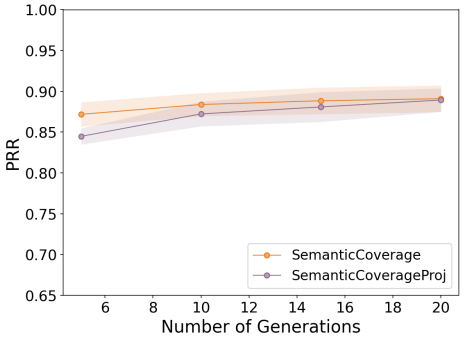


Figure 5: Projection-based coverage and Semantic Coverage converge as the number of generations becomes sufficiently large. SemanticCoverageProj denotes projection-based coverage. Shaded regions indicate mean ±1 std.

C Additional Ablation Studies

Regularization and Eigenspectrum. We next analyze the role of the regularization parameter γ, which controls how strongly small-variance directions are damped. Figure 6 shows that increasing the regularization strength from 0.001 to 1.0 leads to a consistent improvement in PRR, with the best performance attained at $\gamma = 1.0$. The spectral analysis on the 10 largest eigenvalues provides a clear explanation for this behavior. Since centering limits the covariance rank to (N-1), at most nine of these eigenvalues are nonzero for (N=10). The average sorted eigenspectrum in Figure 7b exhibits a sharp drop around 1.0, indicating a separation between a small set of dominant semantic directions and a tail of much smaller eigenvalues. The ECDF in Figure 7a further shows that roughly 65% of the eigenvalues are larger than 1.0. Together, these plots suggest that choosing $\gamma \approx 1.0$ preserves the dominant directions that carry most of the semantic variation, while damping the many weakly informative directions that are more likely to reflect noise. This explains the monotonic improvement observed in the regularization ablation and supports the choice of $\gamma = 1.0$ in the main experiments.

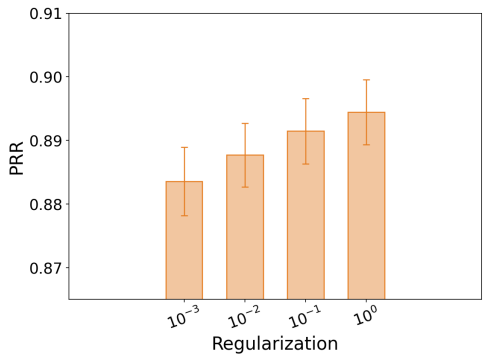
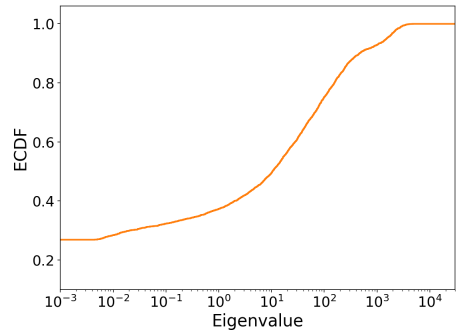


Figure 6: Sensitivity to Regularization for `Qwen3-4B` on CoQA. Capped error bars indicate mean ±1 std.

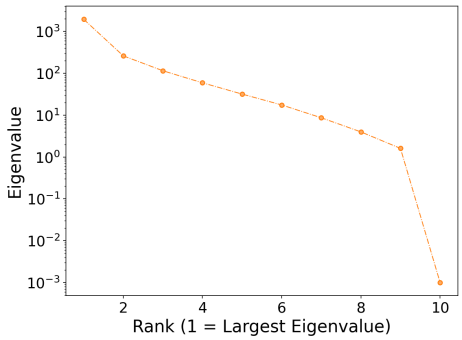
Layers. Figure 8 shows that Semantic Coverage performs best in early-to-middle layers, while performance degrades in later layers. This suggests that the relevant uncertainty signal is strongest before the final layers become too specialized for next-token prediction. The highest PRR is attained in the early layers.

D ROUGE-L Quality Metric

It has been observed that n-gram-based quality metrics such as ROUGE-L (Lin, 2004) often do not reflect the actual quality of generated outputs (Vashurin et al., 2025a). For this reason, in our experiments we use BERTScore, which has been shown to be more robust to surface-form variation than traditional overlap-based metrics (Zhang et al., 2020), and AlignScore, which is sensitive to factual consistency and shows higher correlation with human judgments (Zha et al., 2023). Nevertheless, ROUGE-L is often used in the literature. In Table 5, we report results on the CoQA dataset using



(a) ECDF



(b) Mean Sorted Eigenvalues

Figure 7: Empirical cumulative distribution of eigenvalues and mean sorted eigenvalues for `Qwen3-4B` on CoQA. Top: ECDF. Bottom: Mean sorted eigenvalues.

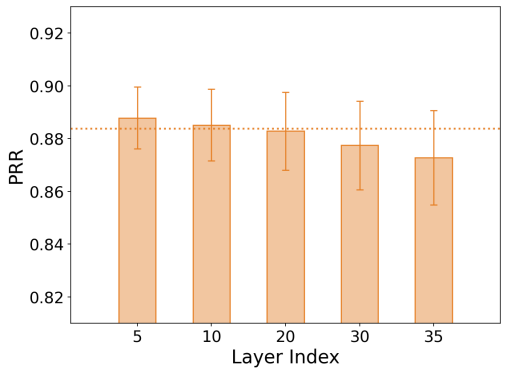


Figure 8: Sensitivity of Semantic Coverage to layer choice for `Qwen3-4B` on the CoQA dataset. The dotted line indicates using the middle layer. Capped error bars indicate mean ±1 std.

ROUGE-L as the quality metric.

E Geometric Interpretation of EigenScore and Semantic Coverage

Chen et al. (2024) introduced EigenScore, which measures dispersion in the sentence-embedding space by using the log-determinant as an approximation to differential entropy. Given a regularized covariance matrix $\tilde{\Sigma}$, EigenScore can be computed as

$$\begin{aligned}\mathrm{eig}(x) &= \frac{1}{d}\log\det(\tilde{\Sigma}) \\ &= \frac{1}{d}\log\prod_{i=1}^{d}\lambda_i \\ &= \frac{1}{d}\sum_{i=1}^{d}\log\lambda_i.\end{aligned}$$

This quantity has a direct geometric interpretation through the covariance ellipsoid associated with $\tilde{\Sigma}$. The eigenvectors define the principal axes of the ellipsoid, while the eigenvalues determine the squared lengths of these axes. Since the volume of the ellipsoid is proportional to $\det(\tilde{\Sigma})^{1/2}$, EigenScore measures the dispersion of sampled representations through their log-volume. Equivalently, it summarizes the geometric mean of the eigenvalues on a logarithmic scale. Focusing specifically on the spectral-summary component, when using the same response-level representations for both methods, Semantic Coverage uses a different summary of the covariance spectrum:

$$\mathrm{cov}(x) = \frac{1}{\mathrm{tr}(\tilde{\Sigma}^{-1})} = \left(\sum_{i=1}^{d}\frac{1}{\lambda_i}\right)^{-1}.$$

This quantity is proportional to the harmonic mean of the eigenvalues and is therefore more sensitive to small eigenvalues. When applied to the same representation geometry the two scores capture different aspects. EigenScore summarizes the overall volume of the covariance ellipsoid, whereas Semantic Coverage is more sensitive to the shape of the ellipsoid, that is, how this volume is distributed across eigendirections. Intuitively, two ellipsoids can have the same volume but different shapes: one may be relatively isotropic, with dispersion spread across several axes, while another may be highly anisotropic, with most of its dispersion concentrated along a small number of axes.

E.1 Mechanism on Synthetic Data

To show why this distinction matters, we consider a simple synthetic experiment in four dimensions (Table 6). Specifically, we fix the generalized variance of a 4×4 covariance matrix at 81 and apply a small regularization value of 0.1 to demonstrate how the different scores respond to the same spectrum. The same qualitative behavior holds when the regularization is small relative to the eigenvalues. As the variance is distributed more uniformly across eigendirections, EigenScore decreases, whereas Semantic Coverage increases. This experiment directly illustrates how the two metrics respond differently to spectral shape, when computed from the same covariance spectrum.

F Generalization to STEM QA

Semantic Coverage has been extensively tested in short-form question answering. Here, we extend the evaluation to more reasoning-heavy QA tasks. In particular, we use the college mathematics, physics, and computer science subsections of the MMLU dataset (Hendrycks et al., 2021). We use accuracy as the quality metric and conduct all experiments with Qwen3-4B. As shown in Table 7, Semantic Coverage maintains strong performance across these subsets.

G Llama Family

In the main paper, we tested two models from the Qwen3 family and one from the Granite3.3 family. To examine how Semantic Coverage generalizes across architectures and training recipes, we add here an evaluation on the CoQA dataset using Llama-3.1-8B (Grattafiori et al., 2024). As shown in Table 8, the findings remain largely unchanged when using Llama-3.1-8B. Semantic Coverage still shows strong performances, achieving the best results when BERTScore is used as the quality metric and the second-best results when AlignScore is used.

H Computational Efficiency Analysis

All experiments use the `Qwen3-4B` model and the hyperparameters and generation settings specified in the experimental setup section. Figure 9 shows the average computational cost, measured as the time each estimator requires to compute the uncertainty score for a single example. Inference time is dominated by the sampling phase, and sampling-based methods exhibit similar overhead. In partic-

PRR

Method	Granite-3.3-8B-it	Qwen3-14B	Qwen3-4B	Mean
Pr(True)	$0.756_{\pm 0.009}$	$0.772_{\pm 0.006}$	$0.743_{\pm 0.004}$	0.757
Perplexity	$0.562_{\pm 0.005}$	$0.725_{\pm 0.009}$	$0.634_{\pm 0.004}$	0.640
LN-Entropy	$0.578_{\pm 0.006}$	$0.720_{\pm 0.011}$	$0.666_{\pm 0.006}$	0.655
CSL	$0.558_{\pm 0.006}$	$0.722_{\pm 0.013}$	$0.635_{\pm 0.007}$	0.638
BoostedProb	$0.555_{\pm 0.004}$	$0.723_{\pm 0.009}$	$0.642_{\pm 0.005}$	0.640
SemanticEntropy	$0.689_{\pm 0.005}$	$0.789_{\pm 0.007}$	$0.780_{\pm 0.006}$	0.753
SemanticDensity	$0.690_{\pm 0.010}$	$0.794_{\pm 0.013}$	$0.735_{\pm 0.012}$	0.740
SentenceSAR	$0.639_{\pm 0.007}$	$0.770_{\pm 0.010}$	$0.740_{\pm 0.006}$	0.716
CoCoAPPL	$0.622_{\pm 0.004}$	$0.766_{\pm 0.008}$	$0.703_{\pm 0.002}$	0.697
AttnScore	$0.655_{\pm 0.013}$	$0.741_{\pm 0.009}$	$0.715_{\pm 0.007}$	0.704
EigenScore	$0.788_{\pm 0.007}$	$0.842_{\pm 0.007}$	$0.813_{\pm 0.007}$	0.814
SemanticCoverage	$\mathbf{0.799}_{\pm 0.005}$	$\mathbf{0.854}_{\pm 0.006}$	$\mathbf{0.832}_{\pm 0.006}$	**0.828**

Table 5: Performance of uncertainty quantification methods using PRR on CoQA across models, with ROUGE-L as the quality metric. Each entry shows mean and standard deviation; the final column reports the average across models.

Eigenvalues	EigenScore	SemanticCoverage
$[81, 1, 1, 1]$	1.170	0.365
$[9, 9, 1, 1]$	1.152	0.491
$[9, 3, 3, 1]$	1.142	0.601
$[3, 3, 3, 3]$	1.131	0.775

Table 6: EigenScore and Semantic Coverage for four covariance spectra with fixed generalized variance of 81. Scores are computed after applying regularization with $\gamma = 0.1$. As the eigenvalues become more uniformly distributed across eigendirections, EigenScore decreases, whereas Semantic Coverage increases.

PRR

Method	Mathematics	Physics	Computer Science	Mean
Pr(True)	0.540	0.749	0.700	0.663
LN-Entropy	0.660	0.649	0.696	0.668
Semantic Entropy	0.683	**0.784**	0.833	0.767
EigenScore	0.715	0.718	0.845	0.759
SemanticCoverage	**0.720**	0.752	**0.849**	**0.774**

Table 7: PRR performance of Pr(True), LN-Entropy, Semantic Entropy, and Semantic Coverage for Qwen3-4B across the college-level Mathematics, Physics, and Computer Science subsets of MMLU. Accuracy is used as the quality metric with PRR.

ular, Semantic Coverage has a slightly lower cost than Semantic Entropy, a lower cost than Semantic Density, and a cost similar to those of EigenScore and LN-Entropy.

I AUROC Evaluation Metric

As our main evaluation metric, we use PRR because it operates directly on continuous quality scores and does not require setting arbitrary correctness thresholds for evaluation. Nevertheless, AUROC is frequently used in the literature. We therefore report complementary results for the

PRR

	CoQA	
Method	BERTScore	AlignScore
Pr(True)	$0.925_{\pm 0.002}$	$0.685_{\pm 0.012}$
Perplexity	$0.923_{\pm 0.002}$	$0.753_{\pm 0.010}$
LN-Entropy	$0.928_{\pm 0.003}$	$0.719_{\pm 0.012}$
G-NLL	$0.953_{\pm 0.002}$	$0.797_{\pm 0.012}$
CSL	$0.917_{\pm 0.003}$	$0.709_{\pm 0.013}$
BoostedProb	$0.936_{\pm 0.003}$	$0.764_{\pm 0.007}$
SemanticEntropy	$0.954_{\pm 0.002}$	$0.803_{\pm 0.005}$
SemanticDensity	$0.941_{\pm 0.002}$	$0.773_{\pm 0.011}$
SentenceSAR	$0.946_{\pm 0.002}$	$0.778_{\pm 0.010}$
CoCoAPPL	$0.951_{\pm 0.002}$	$\mathbf{0.833}_{\pm 0.008}$
AttnScore	$0.932_{\pm 0.002}$	$0.712_{\pm 0.015}$
EigenScore	$0.956_{\pm 0.001}$	$0.794_{\pm 0.008}$
SemanticCoverage	$\mathbf{0.962}_{\pm 0.001}$	$0.818_{\pm 0.007}$

Table 8: Performance of uncertainty quantification methods using the PRR evaluation metric for meta-llama/Llama-3.1-8B on the CoQA benchmark. Each entry shows mean with standard deviation.

Qwen3-4B model using AUROC as the performance metric. As shown in Table 9, the ranking remains mostly unchanged. We also include a small sensitivity study on the correctness threshold used for AUROC in Table 10. Across the different thresholds, Semantic Coverage remains the strongest method. Although the absolute AUROC values vary with the threshold, the ranking is consistent across thresholds.

J Hallucination Detection Case Study

We show several outputs generated by the Qwen3-14B model on the TriviaQA dataset, chosen for its short factoid / entity style answers, with a learned threshold for detecting hallucinations. All hyperparameters are the same as those used in the main body of the paper. We deter-

AUROC

Method	CoQA		Natural Questions		SimpleQA		TriviaQA		Mean
	BERTScore	AlignScore	BERTScore	AlignScore	BERTScore	AlignScore	BERTScore	AlignScore	
Pr(True)	$0.627_{\pm 0.007}$	$0.480_{\pm 0.014}$	$0.530_{\pm 0.005}$	$0.524_{\pm 0.010}$	$0.555_{\pm 0.010}$	$0.457_{\pm 0.035}$	$0.661_{\pm 0.012}$	$0.660_{\pm 0.005}$	0.562
Perplexity	$0.252_{\pm 0.003}$	$0.520_{\pm 0.014}$	$0.319_{\pm 0.006}$	$0.512_{\pm 0.014}$	$0.294_{\pm 0.009}$	$0.509_{\pm 0.008}$	$0.434_{\pm 0.010}$	$0.508_{\pm 0.016}$	0.418
LN-Entropy	$0.333_{\pm 0.008}$	$0.495_{\pm 0.017}$	$0.501_{\pm 0.012}$	$0.522_{\pm 0.023}$	$0.524_{\pm 0.008}$	$0.472_{\pm 0.013}$	$0.442_{\pm 0.007}$	$0.491_{\pm 0.013}$	0.472
CSL	$0.347_{\pm 0.007}$	$0.518_{\pm 0.019}$	$0.445_{\pm 0.008}$	$0.567_{\pm 0.015}$	$0.472_{\pm 0.015}$	$0.550_{\pm 0.020}$	$0.463_{\pm 0.009}$	$0.526_{\pm 0.014}$	0.486
BoostedProb	$0.274_{\pm 0.006}$	$0.554_{\pm 0.012}$	$0.448_{\pm 0.010}$	$0.558_{\pm 0.018}$	$0.550_{\pm 0.012}$	$0.543_{\pm 0.021}$	$0.639_{\pm 0.009}$	$0.698_{\pm 0.012}$	0.533
SemanticEntropy	$0.565_{\pm 0.008}$	$0.618_{\pm 0.011}$	$0.736_{\pm 0.007}$	$0.665_{\pm 0.013}$	$0.745_{\pm 0.011}$	$\mathbf{0.656}_{\pm 0.023}$	$0.723_{\pm 0.012}$	$0.757_{\pm 0.005}$	0.683
SemanticDensity	$0.499_{\pm 0.015}$	$0.573_{\pm 0.004}$	$0.496_{\pm 0.015}$	$\mathbf{0.675}_{\pm 0.015}$	$0.264_{\pm 0.012}$	$0.507_{\pm 0.030}$	$0.721_{\pm 0.011}$	$0.759_{\pm 0.014}$	0.562
SentenceSAR	$0.455_{\pm 0.008}$	$0.565_{\pm 0.016}$	$0.642_{\pm 0.012}$	$0.619_{\pm 0.014}$	$0.654_{\pm 0.006}$	$0.617_{\pm 0.015}$	$0.553_{\pm 0.009}$	$0.590_{\pm 0.013}$	0.587
CoCoAPPL	$0.411_{\pm 0.005}$	$0.643_{\pm 0.018}$	$0.453_{\pm 0.006}$	$0.645_{\pm 0.015}$	$0.338_{\pm 0.008}$	$0.589_{\pm 0.019}$	$0.687_{\pm 0.007}$	$0.753_{\pm 0.008}$	0.565
AttnScore	$0.525_{\pm 0.005}$	$0.492_{\pm 0.016}$	$0.686_{\pm 0.006}$	$0.506_{\pm 0.010}$	$0.628_{\pm 0.009}$	$0.534_{\pm 0.029}$	$0.518_{\pm 0.008}$	$0.509_{\pm 0.006}$	0.550
EigenScore	$0.742_{\pm 0.007}$	$0.603_{\pm 0.009}$	$0.707_{\pm 0.008}$	$0.544_{\pm 0.014}$	$0.751_{\pm 0.015}$	$0.598_{\pm 0.027}$	$0.754_{\pm 0.007}$	$0.743_{\pm 0.006}$	0.680
SemanticCoverage	$\mathbf{0.765}_{\pm 0.010}$	$\mathbf{0.647}_{\pm 0.008}$	$\mathbf{0.752}_{\pm 0.009}$	$0.611_{\pm 0.016}$	$\mathbf{0.807}_{\pm 0.015}$	$0.628_{\pm 0.023}$	$\mathbf{0.800}_{\pm 0.007}$	$\mathbf{0.803}_{\pm 0.006}$	**0.727**

Table 9: Performance of uncertainty quantification methods with the AUROC evaluation metric for `Qwen3-4B` across QA benchmarks. Each entry shows mean and standard deviation; the final column reports the average across datasets and quality measures.

Correctness Measures	BERTScore			AlignScore		
Threshold	0.95	0.90	0.80	0.30	0.50	0.70
Pr(True)	$0.580_{\pm 0.005}$	$0.627_{\pm 0.007}$	$0.731_{\pm 0.031}$	$0.480_{\pm 0.016}$	$0.480_{\pm 0.014}$	$0.490_{\pm 0.015}$
LN-Entropy	$0.355_{\pm 0.010}$	$0.333_{\pm 0.008}$	$0.300_{\pm 0.012}$	$0.500_{\pm 0.016}$	$0.495_{\pm 0.017}$	$0.484_{\pm 0.015}$
SemanticEntropy	$0.579_{\pm 0.004}$	$0.565_{\pm 0.008}$	$0.566_{\pm 0.037}$	$0.624_{\pm 0.007}$	$0.618_{\pm 0.011}$	$0.610_{\pm 0.012}$
SemanticCoverage	$\mathbf{0.764}_{\pm 0.007}$	$\mathbf{0.765}_{\pm 0.010}$	$\mathbf{0.771}_{\pm 0.031}$	$\mathbf{0.647}_{\pm 0.009}$	$\mathbf{0.647}_{\pm 0.008}$	$\mathbf{0.655}_{\pm 0.008}$

Table 10: AUROC evaluation for `Qwen3-4B` on the CoQA dataset under different correctness thresholds for BERTScore and AlignScore.

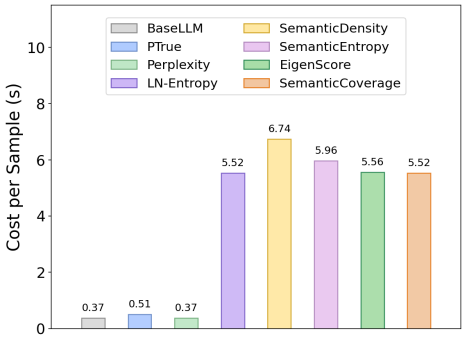


Figure 9: Inference cost comparison of different methods with `Qwen3-4B` on CoQA. BaseLLM denotes the LLM without any uncertainty metric.

mine the optimal threshold for hallucination detection by treating an answer as a hallucination when its BERTScore quality score is lower than 0.9 and by maximizing the G-Mean value (G-Mean $= \sqrt{\mathrm{TPR}\,(1 - \mathrm{FPR})}$). The optimal hallucination thresholds for the different estimators are as follows: Pr(True): 9.944, Perplexity: 3.689, LN-Entropy: 3.491, SemanticEntropy: 8.134, EigenScore: 2.898, and SemanticCoverage: 0.248. A score above the corresponding threshold is classified as a hallucination. Green indicates that the output is not a hallucination, whereas red indicates a hallucinated answer. A ✓ indicates that the output is correctly classified as hallucinated or non-hallucinated, whereas a ✗ indicates that the output is incorrectly classified.

K Artifact Licenses and Intended Use

We use publicly available models and datasets only for research evaluation. Specifically, we evaluate Qwen3-4B, Qwen3-14B, and Granite-3.3-8B-it, and use the QA benchmarks CoQA, SimpleQA, Natural Questions, and TriviaQA. These artifacts are used consistently with their intended research and benchmarking purposes.

Question: What claimed the life of singer Kathleen Ferrier?
Target: Cancer
LLMAns: heart attack
Samples: ['heart attack', 'asthma', 'illness', 'heart failure', 'stroke', 'heart attack', 'heart attack', 'stroke', 'stroke', 'meningitis']
PTrue: 9.922 ✗
Perplexity: 4.673 ✓
LN-Entropy: 7.100 ✓
SemanticEntropy: 17.956 ✓
EigenScore: 1.746 ✗
SemanticCoverage: 0.253 ✓

Question: Name the Danny Boyle biopic about a climber trapped by a boulder for more than five days.
Target: 127 HOURS
LLMAns: 127 hours
Samples: ['127 hours', '127 hours', '127 hours', 'twelve', '127 hours', 'groundhog day', '127 hours', '127 hours', '127 hours', '127 hours']
PTrue: 9.674 ✓
Perplexity: 0.112 ✓
LN-Entropy: 1.416 ✓
SemanticEntropy: 1.897 ✓
EigenScore: 0.571 ✓
SemanticCoverage: 0.166 ✓

Question: Under the Treaty of Breda (31 July 1667), negotiated to end the Second Anglo-Dutch War, the Dutch kept the colony of Suriname conquered from the English, and left the English with what area of land?
Target: New Amsterdam
LLMAns: barbados
Samples: ['New Amsterdam', 'barbadoes', 'the gambia', 'the west indies', 'barbados', 'barbados', 'barbados', 'barbados', 'barbados', 'barbados']
PTrue: 10.450 ✓
Perplexity: 0.285 ✗
LN-Entropy: 2.592 ✗
SemanticEntropy: 5.065 ✗
EigenScore: 2.152 ✗
SemanticCoverage: 0.200 ✗

Question: In the context of UK government, for what does the C stand in the acronym COBRA?
Target: Cabinet
LLMAns: committee
Samples: ['commissioner', 'committee', 'committee', 'county', 'commissioners', 'central', 'chancellor of the exchequer', 'crisis', 'commissioner', 'conveyancing']
PTrue: 9.962 ✓
Perplexity: 6.223 ✓
LN-Entropy: 4.769 ✓
SemanticEntropy: 7.773 ✗
EigenScore: 8.686 ✓
SemanticCoverage: 0.973 ✓

Question: Measuring from the closest point of each of these countries to the equator, which is the furthest from the equator?
Target: Canada
LLMAns: india
Samples: ['india', 'ecuador', 'portugal', 'france', 'portugal', 'chile', 'england', 'india', 'sweden', 'france']
PTrue: 9.955 ✓
Perplexity: 5.154 ✓
LN-Entropy: 6.874 ✓
SemanticEntropy: 20.757 ✓
EigenScore: 3.778 ✓
SemanticCoverage: 0.333 ✓

Question: Who wrote the nonsense poem "Jabberwocky"?
Target: Lewis Carroll
LLMAns: lewis carroll
Samples: ['lewis carroll', 'lewis carroll', 'lewis carroll', 'lewis carroll', 'lewis carroll', 'lewis carroll', 'lewis caroll', 'lewis carroll', 'carroll', 'lewis carroll']
PTrue: 9.430 ✓
Perplexity: 0.058 ✓
LN-Entropy: 1.865 ✓
SemanticEntropy: 0.740 ✓
EigenScore: 0.808 ✓
SemanticCoverage: 0.166 ✓

Question: Which Scottish actress played Catherine of Aragon in the 1970 TV series, "The Six Wives of Henry VIII", but is much better known for her role as Margaret Meldew in "One Foot in the Grave"?
Target: Annette Crosbie
LLMAns: joan collins
Samples: ['joan plowright', 'joan collins', 'marion', 'jan francis', 'joan collins', 'marjorie proops', 'joanna lumsden', 'anastasia hewetson', 'shirley steward', 'emily mortimer']
PTrue: 11.705 ✓
Perplexity: 0.417 ✗
LN-Entropy: 2.124 ✗
SemanticEntropy: 8.938 ✓
EigenScore: 7.182 ✓
SemanticCoverage: 0.498 ✓